White Paper Report

Report ID: 102670

Application Number: HD5132911

Project Director: Micki McGee (mmcgee@fordham.edu)

Institution: Fordham University

Reporting Period: 4/1/2011-5/31/2012

Report Due: 8/31/2012

Date Submitted: 8/31/2012

White Paper

Grant # HD-51329-11

Compatible Databases Initiative: Toward Interoperable Data for Network Mapping and Visualization

Micki McGee (Fordham University), Principal Investigator
Richard Edwards (Ball State University)
Cristina Pattuelli (Pratt School of Information Science)
Edward Whitley (Lehigh University)

Fordham University

August 31, 2012

A. Narrative Description

The Compatible Databases Initiative, launched in May 2011 with support from the National Endowment for the Humanities, set out to facilitate a conversation about interoperable data among humanities scholars working in various disciplines who have been involved in building person-centric (or prosopographic) datasets and relational databases as tools for historical and cultural research and, more recently, for the purpose of rendering network visualizations and analyses.

The initial cultural network mapping and database projects identified as part of the initiative included: *The Crowded Page* (Andrew Jewell and Edward Whitley), *Phylo* (Chris Alen Sula and David Morrow), the *Yaddo Archive Project* (Micki McGee and Richard Edwards), *Explore Thomas Cole* (Charles Forcey/Historicus, Inc.), and the *Orlando Project* (Patricia Clements, Susan Brown, and Isabel Grundy.)[1] As the project unfolded we learned of the *Linked Jazz* project (Cristina Pattuelli), and have invited Pattuelli's participation.

The problem that has emerged across these projects is that the unique data structures in these projects lack interoperability — that is, data collected about any given historical figure in one database cannot be readily integrated with data from another project. No open standards or best practices for database architecture for humanities scholarship have yet been widely adopted. While archivists have initiated the move toward a standardized data formats (specifically with the development of EAD-CPF[2]), for the most part scholars in the humanities and social sciences have continued to work independently, creating databases that, while innovative, are isolated and incompatible from one project to another.

The importance of fostering data interoperability now cannot be overstated. As network data visualization tools become more widely available, more and more scholars will be developing datasets for visualizations. Fostering compatibility among datasets creates the possibility of extending the maps of cultural communities across domains (literature, philosophy, anthropology, art history and criticism, sociology), and allows scholars to mine deeply into the data that underlie networks of social relationships and ideas. In fact, these data constitute a potentially amazing window on our intellectual and cultural history – if they are interoperable and linked.

Making this networked data compatible will enable a rich social and intellectual cartography that will be of value across humanities disciplines, especially to historians, literary theorists, and philosophers, as well as to general audiences of individuals concerned with the history of

---

[1] *The Crowded Page* (www.crowdedpage.org); *Phylo* (http://www.newmedialab.cuny.edu/phylo); *Yaddo Archive Project* (preliminary visualization) www.yaddo-circles.org; *Explore Thomas Cole* (http://www.explorethomascole.org/); *Orlando* (http://www.ualberta.ca/ORLANDO); and *Linked Jazz* (www.linkedjazz.org)

[2] See EAC-CPF is maintained by the Society of American Archivists in partnership with the Berlin State. n.d. "Society of American Archivists and the Berlin State Library." Retrieved August 20, 2012 (http://eac.staatsbibliothek-berlin.de/).

ideas. Establishing and disseminating guidelines for best practices in data collection and architecture will ensure that as multiple databases are developed that their contents are interoperable. This year long project initiated a conversation on developing and promoting the use of data standards for interoperability. The details of our accomplishments are discussed below.

B. Project Activities and Recommendations

The primary activity of the Compatible Databases Initiative was a two-day conference in New York City (September 23-25, 2011) held at Fordham University and the New York Public Library. This conference brought together the leaders of the aforementioned projects along with other data architecture and data visualization experts, technologist and programmers, and humanities scholars. (See *Appendix 1. Meeting Participants*). Internationally renowned data visualization expert Katy Börner delivered a Friday evening keynote address, and EAD-CPR instigator and SNAC Project Director Daniel Pitti presented a Saturday morning keynote. These lectures were followed by a day and half of intensive project presentations and conversations on key issues in data interoperability for humanities scholarship and network analysis.

The outcomes of this meeting include: 1) an outline of issues in the development of interoperable data standards; 2) best practices recommendations for researchers currently working with person-centric databases; and 3) a recommendation that two of the projects (*The Crowded Page* and the *Yaddo Archive Project*) consider a pilot project merging their data. We have since extended this planning to include the innovative *Linked Jazz* project. (Please see "Section E. Continuation of the Project" for more on this planning.) Additional conversations have resulted in the recommendation that our work be tied more closely to the LODLAM (Linked Open Data in Libraries, Archives, and Museums).

1) Outline of Issues from the Compatible Databases Initiative Conference

During the initial two-day conference in New York City, the participants prioritized the major issues involved in developing guidelines and standards for interoperable data. From a list of issues generated by the entire working group, each participant voted for their top three issues, resulting in the following working list of the top ten issues in our compatible data conversation. (Note: The following list is arranged by highest vote tallies.)

a) Establish new standards of evidence in the creation and maintenance of person-centric datasets
b) Develop methods and guidelines for linking person-centric data back to their primary data sources
c) Create public APIs for use in the creation of interoperable datasets
d) Address the relationship between controlled and uncontrolled vocabularies in person-centric datasets (authoritative institutional vs. folksomic vocabularies)
e) Consider and address how the standardization of interoperable person-centric datasets will affect existing and emerging visualization techniques used for scholarly inquiry
f) Develop standardized approaches for dealing with exceptions and "uncertain data" in

person-centric datasets

g) Account for "platform instability" and the pace of technological change in designing standards for interoperable data
h) Include Resource Description Frameworks (RDFs) and semantic web issues that are involved with interoperable datasets
i) Develop standards that will maintain local scholarly control over different datasets and primary data sources
j) Consider and address the interoperability problems that will arise around different data qualities, ontologies, and confidence standards across different datasets

An idea that emerged from Compatible Databases Initiative meeting, in part because of the vital participation of information scientist Katy Börner, was the notion that various science disciplines that have grappled with data interoperability may have models that could be adapted to humanities queries. All the research questions and methods of inquiry in the sciences and the humanities differ greatly, there may still be significant parallels that can inform our work on the issues listed above. To this end, we are considering proposing the development of a Summer Institute or workshop to bring together scientists and humanities scholars to consider how the principles of interoperability used in the sciences might be adapted for humanities inquiry.

2) Best practices recommendations for researchers currently working with person-centric relational databases:

a) Conforming personal and corporate or entity names to Library of Congress Authorities names, and, as the National Archival Authorities Cooperative begins establishing authoritative guidelines, to conform to these controlled vocabularies for naming
b) Establishing "alternate names" tables in SQL databases (or in additional columns in spreadsheet data collection) to ensure the possibility of cross-referencing
c) For relationship codes, we recommend the use of the controlled vocabularies developed by Ian Davis and Eric Vitiello, Jr. for RDF/XML, JSON, and TURTLE available at http://vocab.org/relationship. In addition, the expanded vocabulary developed in *The Crowded Page* White Paper[3] serves as a model for how to adapt and expand the baseline ontology from vocab.org.
d) When idiosyncratic relationship information is required by a particular research topic (for example, the *Yaddo Archive Project* contains a vast number of recommendation letters for which authors and subjects are linked as "RecommendedBy") then an additional table that includes these non-standard predicate code "RecommendedBy" should be used (or, in the case of spreadsheet data capture, an additional column would be required to capture these data).

3) Fostering the use of linked open data among scholars:

Some skepticism has been expressed that humanities scholars will be willing to conform to

[3] Jewell, Andrew and Edward Whitley (2011), "White Paper Grant #HD5044008 The Crowded Page," Available online: https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-50440-08 (Last accessed: August 30, 2012).

data standards that include controlled vocabularies that are not tailored to their specific research questions. Commenting at National Archival Authorities Cooperative planning meeting at the National Archive (May 21, 2012) archivist Mark A. Matienzo noted (regarding individual scholars' data): "what you do in the privacy of your own database is your own business." This view is understandable, as it has been the practice up until now. However a traditional view of the humanities scholar laboring alone on his or her monograph may well be increasingly anachronistic. Certainly scholars working in the humanities, particularly in the digital humanities, have grown interested in creating data that can be repurposed in compatible formats.

Approaches to fostering interoperable data include:

a) Data standard mandates: One approach to ensuring interoperability would involve mandating the use of particular data standards for federally- or state-funded projects and as a condition of library acceptance for archiving of databases. In fact, such standards could be incorporated into the open data standards being developed by the Endowment.

b) Tools and APIs to facilitate interoperability: Another more flexible (and perhaps appealing) approach would be the development of a research and data capture tool for scholars that would:

   i. Automate the process of conforming (and disambiguation of names for persons, corporate bodies and other entities)
   ii. Allow for easy multiple tagging for relationship types to generate both controlled vocabulary relationships and the idiosyncratic relationship typologies required by particular research questions. *Project RoSE* (http://rose.english.ucsb.edu) directed by Alan Liu, has begun these explorations.

   Ideally, such a data capture tool would be developed to operate as a plug-in for already existing open access bibliographic software such as Zotero. (See "Section E. Continuation of the Project" for additional discussion.)

c) Dissemination of interoperable data goals at professional conferences: To disseminate these ideas, we recommend developing panel presentations and workshops for the key humanities conferences on the principles and practices of linked open data for scholars. Conferences to be targeted for such panels would be the annual meetings of the American Historical Association, the American Sociological Association, the American Studies Association, and the Modern Language Association, along with the international Digital Humanities Conference.

d) Training the next generation of humanities scholars in interoperable data practices: Although engaging humanities scholars currently at work on developing historical person-centric databases is important, ultimately the success of this initiative rests with educating an emerging generation of humanities scholars in the principles and practices of data interoperability. To this end, we recommend that digital humanities graduate programs (and indeed all graduate humanities programs) include principles of linked open data and data interoperability in their graduate methods courses.

Toward this end, we recommend the development of graduate level courses on linked open data for the traditional humanities disciplines to be piloted in graduate programs such as the Praxis Program being developed at the University of Virginia's Scholar's Lab and other programs nationwide.

e) Link scholars to the LODLAM community: Energetic expansion of the Compatible Databases Initiative will be achieved by more closely aligning the project with the Linked Open Data in Libraries, Museums, and Archives movement that was also launched in 2011 with NEH support. To this end, we are considering adopting the name suggested by Elton Barker (Department of Classics, Open University): LODS (Linked Open Data for Scholars), and encouraging the LODLAM group to include "S" for scholars in their emerging dialogues, as LODSLAM (Linked Open Data for Scholars, Libraries, Archives, and Museums) or LODLAMS (Linked Open Data for Libraries, Archives, Museums and Scholars).

C. Audiences

Although the immediate audience for the Compatible Databases Initiative is focused on scholars working with prosopographic research in the digital humanities (including the disciplinary specializations of art history, classics, history, literature, music history and American Studies and other area studies specialties) the impact of this research will be far-reaching in that the development and implementation of interoperable data standards will allow for the creations of discovery and reference tools for general audiences allowing them to map the links between historical persons, entities, corporate and institutional bodies.

D. Evaluation

Evaluation of the Compatible Databases Initiative will utilize open peer review and public commentary on the resulting white paper. The white paper will be hosted by MediaCommons at their main website using the CommentPress software. (Further information about this digital scholarly community and its practice of hosting the open peer review of white papers can be found at http://mediacommons.futureofthebook.org/).

This open and public process will invite comments and responses from members of (a) the full Compatible Data Group, and (b) the worldwide Linked Open Data community to encourage further discussion, debate, and sharing of the key issues around interoperable data standards. This is also a method for further dissemination of the group's findings and suggestions.

E. Continuation of the Project

Five activities for project continuation have been identified:

1) Develop closer links to the Linked Open Data in Libraries, Museums, and Archives project that was also launched in 2011 with NEH support. Although the Principal Investigator on the Compatible Databases Initiative took part in the June 2011 LOD-LAM Summit in San Francisco, and at several subsequent LODLAM meetings in New York City, more and stronger links to the LODLAM community will be key to the

project's success.

2) To demonstrate the value of interoperable data, three projects (*The Crowded Page*, *Linked Jazz*, and *Yaddo Circles*) will seek support to: a) conform their data to explore overlaps between the communities included in their datasets in online network visualization and b) develop the research and data capture tool described above.
3) Explore the possibility of conducting presentations at professional association meetings in the coming year, with the MLA, the AHA, and the American Studies Association identified as the first three conferences for this outreach.
4) Foster development of a data standards course for graduate level humanities disciplines will be discussed with leaders of new digital humanities graduate certificate programs, and with other humanities graduate programs.
5) Consider planning for a Summer Institute or workshop to bring together scientists and humanities scholars to consider how the principles of interoperability used in the sciences might be adapted for humanities inquiry.

F. Long-Term Impact

By fostering the goal of data interoperability among humanities scholars engaged in person-centric research, the Compatible Databases Initiative contributes to the long-term goal of supporting linked open data for among humanities researchers. This critical technical standards work lays the ground work for general research and discovery tools to map the relationships between persons, institutions, and the spread of ideas that will be invaluable to not only humanities scholars, but to the general public.

G. Works Cited

Davis, Ian and Vitiello, Eric, Jr. (2004, last updated 2010, April 19). RELATIONSHIP: A vocabulary for describing relationships between people. Available online: http://vocab.org/relationship/ (Last accessed: August 30, 2012).

Jewell, Andrew and Edward Whitley (2011). "White Paper Grant #HD5044008 The Crowded Page," Available online: https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-50440-08 (Last accessed: August 30, 2012).

Pattuelli, M. Cristina. (2011). Mapping people-centered properties for Linked Open Data. *Knowledge Organization, 38* (4), 352-359.

H. Appendices

Appendix 1. Compatible Databases Meeting Participants
Appendix 2. Related Projects and Resources

Appendix 1. Meeting Participants, September 23-25, 2011

Katy Börner (Indiana University/Bloomington)
Susan Brown (University of Guelph)
Terry Capatano (Columbia University)
Elizabeth Cornell (Fordham University)
Craig Dietrich (University of Southern California) *
Richard Edwards (Ball State University)
Charles Forcey (Historicus, Inc.)
Daniel Pitti (University of Virginia/IATH)
Jon Ippolito (University of Maine) *
Alan Liu (University of California at Santa Barbara) *
Micki McGee (Fordham University)
Aditi Muralidharan (University of California, Berkeley)
Asik Prahan (Indiana University/Bloomington)
Chris Alen Sula (Pratt Institute, School of Library and Information Science)
William Stingone (New York Public Library, Division of Manuscripts and Archives)
Robert Weidman (Lehigh University)
Edward Whitley (Lehigh University)

* participated via teleconference.

Appendix 2. Related Projects and Resources

Compatible Data Initiative
www.compdb.blogspot.com

The Crowded Page
www.crowdedpage.org

Linked Jazz
www.linkedjazz.org

Linking Lives
http://archiveshub.ac.uk/linkinglives

Linked Open Data for Libraries, Museums, and Archives
http://lod-lam.net
http://lodlam.net

Person Data Repository
http://pdr.bbaw.de/english

Phylo
http://phylo.info

Project RoSE: A Research Oriented Social Environment
http://rose.english.ucsb.edu

Scalar
http://scalar.usc.edu

Sci2 Tool
https://sci2.cns.iu.edu/user/index.php

Social Networks and Archival Context Project
http://socialarchive.iath.virginia.edu

ThoughtMesh
http://thoughtmesh.net